Journal of Theoretical Biology 353 (2014) 19–23

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

A protein structural classes prediction method based on PSI-BLAST profile

Shuyan Ding [a,*], Shoujiang Yan [b], Shuhua Qi [a], Yan Li [b], Yuhua Yao [b,**]

[a] Department of Sciences, Dalian Nationalities University, Dalian, Liaoning 116600, PR China
[b] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou, Zhejiang 310018, PR China

AUTHOR-HIGHLIGHTS

- The long-range information is extracted.
- The linear correlation coefficient is used to extract information from PSSM.
- The stand-alone version of our method (LCC-PSSM) is constructed.

ARTICLE INFO

Article history:
Received 16 December 2013
Received in revised form
27 January 2014
Accepted 24 February 2014
Available online 4 March 2014

Keywords:
Feature selection
Support vector machine
Position-specific scoring matrix

ABSTRACT

Knowledge of protein structural classes plays an important role in understanding protein folding patterns. Prediction of protein structural class based solely on sequence data remains to be a challenging problem. In this study, we extract the long-range correlation information and linear correlation information from position-specific score matrix (PSSM). A total of 3600 features are extracted, then, 278 features are selected by a filter feature selection method based on 1189 dataset. To verify the performance of our method (named by LCC–PSSM), jackknife tests are performed on three widely used low similarity benchmark datasets. Comparison of our results with the existing methods shows that our method provides the favorable performance for protein structural class prediction. Stand-alone version of the proposed method (LCC–PSSM) is written in MATLAB language and it can be downloaded from http://bioinfo.zstu.edu.cn/LCC-PSSM/.

1. Introduction

Knowledge of structural class information of a given protein plays an important role in the prediction of secondary structure, tertiary structure and function analysis from the amino acid sequence (Anand et al., 2008). Levitt and Chothia (1976) studied the polypeptide chain topologies in a dataset of 31 globular proteins and categorized the protein domains of known structure into four structural classes: all-α, all-β, α/β and $\alpha+\beta$ classes. With the rapid development of sequencing technology, the exponential growth of newly discovered protein sequences by different scientific communities has made a large gap between the number of sequence-known and the number of structure-known proteins. Hence, there exists a critical challenge to develop automated methods for fast and accurate determination of the protein structural classes in order to reduce the gap.

During the past two decades, prediction of protein structural class based on the amino acid sequence became a hot topic and many different methods have been proposed to address this problem. There are generally two aspects: sequence feature extraction and classification algorithm. Various sequence features have been applied to represent protein sequences, including amino acid composition (AAC) (Nakashima et al., 1986; Zhou, 1998), pseudo amino acid composition (PseAA) (Chen et al., 2012a; Chou, 2001; Ding et al., 2007; Li et al., 2009; Liao et al., 2012; Qin et al., 2012; Sahu and Panda, 2010; Wu et al., 2010; Xiao et al., 2006, 2008a, 2008b; Zhang and Ding, 2007; Zhang et al., 2008), polypeptide composition (Luo et al., 2002; Sun and Huang, 2006), functional domain composition (Chou and Cai, 2004; Sommer et al., 2004), PSI-BLAST profile (Chen et al., 2008; Liu et al., 2010, 2012), and predicted secondary structure information (Ding et al., 2012; Kurgan et al., 2008a, 2008b; Mizianty and Kurgan, 2009; Yang et al., 2010). Meanwhile, many machine learning algorithms have been already used to implement the protein structural class predictions, such as neural network (Cai and Zhou, 2000), support vector machine (SVM) (Anand et al., 2008; Cai et al., 2001, 2002; Chen et al., 2006; Zhang

* Corresponding author.
** Corresponding author.
E-mail addresses: sunnyday1979@163.com (S. Ding), yaoyuhua2288@163.com (Y. Yao).

http://dx.doi.org/10.1016/j.jtbi.2014.02.034
0022-5193

et al., 2012), fuzzy clustering (Shen et al., 2005), Bayesian classification (Wang and Yuan, 2000), and rough sets (Cao et al., 2006).

Among the above sequence feature extraction methods, features extracted from the predicted secondary structure sequence and PSI-BLAST profile rather than directly from the amino acid sequence itself can present a higher prediction accuracy (Chen et al., 2008; Liu et al., 2010, 2012; Yang et al., 2010; Kurgan et al., 2008a, 2008b; Mizianty and Kurgan, 2009; Zhang et al., 2012). Usually, with the addition of predicted protein secondary structure, the features extracted from predicted secondary structure sequence can provide the higher overall accuracy than other methods. However, the trade-off is that these methods must run a secondary-structure predictor to generate their input, which is somehow more demanding computationally. Features extracted from PSI-BLAST profile can provide more evolutionary information, which can also provide the favorable prediction results.

In this study, we try to extract more evolutionary information solely from the PSI-BLAST profile to further improve the prediction accuracy. A feature set consisting of 278 features is constructed by feature selection method based on 1189 dataset. Jackknife tests on the low-similarity datasets show that the current method presents satisfying prediction accuracies in comparison with the existing methods.

As demonstrated by a series of recent publications (Chen et al., 2012c, 2013; Min et al., 2013; Xiao et al., 2013a; Xu et al., 2013a, 2013b), and summarized in a comprehensive review (Chou, 2011), to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (iii) introduce or develop a powerful algorithm to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps.

2. Materials and methods

2.1. Materials

A total of three low-similarity datasets were used to design and test the new method. The 1189 dataset includes 1092 protein domains with sequence similarity lower than 40%, which consists of 223 all-α class proteins, 294 all-β class proteins, 334 α/β class proteins, and 241 $\alpha+\beta$ class proteins (Wang and Yuan, 2000). The 25PDB dataset includes 1673 protein domains with sequence similarity lower than 25% of which 443 are all-α class proteins, 443 are all-β class proteins, 346 are α/β class proteins and 441 are $\alpha+\beta$ class proteins (Kurgan and Homaeian, 2006). The third protein dataset, referred to as 640, was first studied in Chen et al. (2008). It contains 640 proteins with 25% sequence identity of which 138 are all-α class proteins, 154 are all-β class proteins, 177 are α/β class proteins and 171 are $\alpha+\beta$ class proteins.

2.2. Feature extraction

In order to improve the prediction accuracy of low-similarity proteins, we extract the evolutionary information from PSI-BLAST profile which is represented as a so-called position-specific score matrix (PSSM). The features are extracted based on 1189 dataset.

PSI-BLAST is a tool that produces PSSM constructed from a multiple alignment of the highest scoring hits in an initial BLAST search. We use each protein sequence (called query sequence) as a seed to search and align homogeneous sequences from NCBI's NR database (ftp://ftp.ncbi.nih.gov/blast/db/nr) using the PSI-BLAST program (Altschul et al., 1997) with three iterations and a cutoff E-value 0.001. PSSM is a log-odds matrix of size $L \times 20$, where L is the length of the query amino acid sequence and 20 is due to the 20 amino acids. The (i,j)th entry of the matrix represents the score of the amino acid in the ith position of the query sequence being mutated to amino acid type j during the evolution process.

In this study, the PSSM elements are scaled to the range from 0 to 1 using the following sigmoid function:

$$f(x) = 1/(1+e^{-x})$$

where x is the original PSSM value.

For convenience, let us denote

$$D = (P_1, P_2, \ldots, P_{20})$$

as the PSSM of the query sequence S with length L, where, for example,

$$P_j = (p_{1,j}, p_{2,j}, \ldots, p_{L,j})^T$$

T is the transpose operator, and $p_{i,j}$ ($i=1, 2,\ldots, L$) denotes the score of the amino acid in the ith position of S being mutated to the jth amino acid during the evolution process.

To successfully use support vector machine (SVM) as a powerful classifier, the key is how to effectively define a feature vector to formulate the statistical samples concerned. According to Eq. (6) of Chou (2011), the feature vector for any protein, peptide, or biological sequence is none but a general form of pseudo amino acid composition or PseAA (Chou, 2001) that can be formulated as

$$P = (\Psi_0, \Psi_1, \ldots, \Psi_g, \ldots, \Psi_G)^T \quad (1)$$

where T is a transpose operator, the component $\Psi_g (g = 0, 1, \ldots, G)$ is a vector which depends on how to extract the desired information from the statistical samples concerned.

The linear correlation coefficient, which is also called Pearson's r, is the most widely used measure of the association between pairs of values. In this paper, we combine the long-range correlation information and the linear correlation information of P_s and $P_t (s \neq t)$ together to perform the feature extraction. In order to realize this idea, the linear correlation coefficient of $(p_{1,s}, p_{2,s}, \ldots, p_{L-g,s})^T$ and $(p_{g+1,t}, p_{g+2,t}, \ldots, p_{L,t})^T$ is used to reflect the average correlation between two residues separated by a gap of g along the sequence S.

For convenience, we denote

$$A_{s,t,g} = \frac{1}{L-g} \sum_{i=1}^{L-g} p_{i,s} \times p_{i+g,t} \quad (2)$$

$$B_{s,g} = \frac{1}{L-g} \sum_{i=1}^{L-g} p_{i,s} \quad (3)$$

$$C_{t,g} = \frac{1}{L-g} \sum_{i=g+1}^{L} p_{i,t} \quad (4)$$

$$D_{s,g} = \frac{1}{L-g} \sum_{i=1}^{L-g} p_{i,s}^2 - \left(\frac{1}{L-g} \sum_{i=1}^{L-g} p_{i,s} \right)^2 \quad (5)$$

$$E_{t,g} = \frac{1}{L-g} \sum_{i=g+1}^{L} p_{i,t}^2 - \left(\frac{1}{L-g} \sum_{i=g+1}^{L} p_{i,t} \right)^2 \quad (6)$$

Then, we define

$$LCC_{s,t,g} = (A_{s,t,g} - B_{s,g} \times C_{t,g}) / \sqrt{D_{s,g} \times E_{t,g}} \quad (7)$$